Opportunistic Bits in Short-Packet Communications: A Finite Blocklength Perspective

Mingxi Yin, Yuli Yang, *Senior Member*, *IEEE*,
Jen-Ming Wu, *Member*, *IEEE*, and Bingli Jiao, *Senior Member*, *IEEE*

***Abstract*—In this paper, the concept of opportunistic bits (OBs) is developed in short-packet communications and investigated from a finite blocklength perspective. In the OB-based transmission, the data unit of a packet is divided into two parts: OBs and conventional bits (CBs). The OBs are not physically transmitted but used to indicate the index of the time slot (TS) when the packet containing CBs is transmitted. The loading of a bulk of OB-based packets into multiple TSs can be modelled as a Repeated Balls-into-Bins process with a multi-queue storage. If the bulk is not large enough, certain combination(s) of OBs will not appear, which leaves certain TS(s) empty and hence reduces the TS load efficiency. To evaluate the OB-based transmission performance, we formulate its maximal payload rate and TS load efficiency. With the aid of these two formulations, the energy gain, the goodput, and the latency of OB-based short-packet communications are derived and obtained in analytical forms. For achieving further insights, illustrative numerical results on the resource utilisation efficiency and the performance not only substantiate the advantages of the OB-based transmission over the conventional but also provide useful tools and specifications for its design in massive short-packet communications.**

***Index Terms*—Opportunistic bit (OB), short-packet communications, finite blocklength, time slot (TS) load efficiency, maximal payload rate, energy gain, goodput.**

I. Introduction

To accommodate wireless services with low latency and ultra-high reliability in future mobile networks, e.g., massive access in the Internet of Things (IoT), short-packet communications become popular solutions to support mission-critical applications [1], [2]. Advanced wireless technologies, e.g., multi-antenna configurations [3], full-duplex transmissions [4], non-orthogonal multiple access [5], optimisation of resource allocation [6], and cross-layer design [7], have been further explored or redeveloped in short-packet communications.

As shown in [8] and [9], transmission protocols and techniques designed for long packets in conventional cellular networks are not suitable for short packets in the IoT networks. Moreover, Shannon's classical analysis framework, i.e., the optimal coding rate converging to channel capacity [10], is invalid for short-packet communications [11], [12] and, therefore, the maximal coding rate of short packets has been investigated in the finite blocklength regime [13], [14]. Later on, simple asymptotic approximations of converse and achievability bounds on the maximal coding rates, particularly useful for short packets, have been provided in [15]. Further, close bounds on the maximum coding rates have been studied in wireless channels of practical interest, e.g., no *a priori* channel state information available and pilot-assisted or noncoherent transmissions utilized in Rayleigh and Rician fading [16], [17].

These information-theoretic advances have established a basis for the design of short-packet communications. Compared with its long-packet counterpart, short-packet communications bring about new communication models and associated challenges that need to be addressed. Firstly, the IoT device configuration and the service latency requirement limit the transceiver's buffer size [18], [19]. Secondly, the meta-data size, arising from control information, is comparable to the payload size in short packets [20], [21].

Recently, the philosophy of permutation modulation [22], [23] has been embraced by upper layer(s) to increase the goodput in a straightforward way [24]–[26]. Specifically, the concept of opportunistic bits (OBs) was proposed in [24] to increase the achievable data rate for point-to-point transmissions of massive transmission control protocol/Internet protocol (TCP/IP) packets. OBs are a portion of information bits in the data unit (DU) of a packet, which are used to indicate the index of the time slot (TS) when the packet is transmitted. Since OBs are not involved in the physical transmissions of packets, the energy and spectral efficiency of the OB-based communication systems will be improved compared with conventional ones. The information bits other than OBs in the DU of a packet are referred to as conventional bits (CBs), which are physically transmitted in a conventional way. In other words, the DU of an OB-based packet contains CBs only and the packet is transmitted in the TS mapped by OBs. At the receiver, the received packets will be reordered according to their sequence numbers prescribed by the TCP protocol. In [27], this OB transmission strategy was generalised into networks to pair packets of the same OBs and hence realize end-to-end transmissions.

Apparently, the same OB size will take a larger percentage in a smaller DU as the blocklength decreases. Therefore, the OB-based design will achieve higher resource efficiency in short-packet communications than in their longer-packet counterparts. Motivated by this, we develop the OB paradigm for short-packet communications in this work. Specifically, the maximal payload rate of OB-based short-packet communications is formulated in the finite blocklength regime to remove

This work was supported in part by National Key Research & Development Program of China under Grants 2020YFB1807802, 2019YFE0113200, 2018YFB2202202, 2016ZX03001018-005, and National Natural Science Foundation of China under Grant 61771188. The calculations were supported by the High-performance Computing Platform of Peking University.

M. Yin and B. Jiao are with the Department of Electronics, Peking University, Beijing 100871, China (e-mail: yinmx@pku.edu.cn, jiaobl@pku.edu.cn).

Y. Yang (*corresponding author*) is with the School of Engineering, University of Lincoln, Lincoln LN6 7TS, U.K. (e-mail: yyang@lincoln.ac.uk).

J.-M. Wu is with the Institute of Communications Engineering, National Tsing Hua University, Hsinchu 30013, Taiwan (e-mail: jmwu@ee.nthu.edu.tw).

the ideal assumptions of infinite blocklength and arbitrarily small error probability from information-theoretic analysis.

However, due to the limitation on a transceiver's buffer size, certain combination(s) of OBs will be missing if the bulk of packets in a short-packet communication is not large enough. Therefore, certain TS(s) will be empty and the TS load efficiency will be affected, which is a main concern in OB-based short-packet communications. To formulate the TS load efficiency in the scenario of limited buffer size, we model the loading of a bulk of OB-based packets into multiple TSs as a Repeated Balls-into-Bins (RBB) process with a multi-queue storage. With the aid of this modelling, the TS load efficiency is achieved to verify the feasibility of the OB concept embraced by short-packet communications.

Based on the formulations of maximal payload rate and TS load efficiency, the metrics of energy gain, goodput and latency are investigated to further evaluate the resource utilisation efficiency and performance of the OB-based short-packet communications. In particular, our main contributions in this paper are three-fold:

- To further benefit from the high resource efficiency of OB-based transmissions, the OB concept is utilized in short-packet communications and its maximal payload rate is formulated in the finite blocklength regime.
- To provide an analysis framework for practical applications, the TS load efficiency of OB-based short-packets is formulated to validate the utilization.
- To evaluate the performance of OB-based short-packet communications, their energy gain, goodput, and latency are achieved in analytical forms. Moreover, illustrative numerical results on these metrics are demonstrated to gain further insights.

The remainder of this paper is organized as follows. Firstly, the short packet structure and the OB concept are introduced in Section II. Subsequently, the maximal payload rate and the TS load efficiency of OB-based short-packet communications are formulated in Sections III and IV, respectively. Based on these two formulations, Section V evaluates the resource utilisation efficiency and performance of the OB concept utilized for short packets in the metrics of energy gain, goodput and latency. Finally, this paper is concluded in Section VI.

Throughout this paper, the following mathematical notations are used: Boldface uppercase and lowercase letters denote matrices and vectors, respectively. The number of entries in the vector $\mathbf{x}$ that are equal to y is denoted by $\mathrm{num}(\mathbf{x}, y)$. Moreover, $\mathcal{D}(\cdot)$ and $\mathcal{B}(\cdot)$ denote the binary-to-decimal and decimal-to-binary converters, respectively. In addition, $\mathbb{P}[\cdot]$ denotes the probability of an event and $\mathbb{E}[\cdot]$ represents the expectation (mean) operator.

II. System Model

In this section, the short packet structure is presented and the OB concept is utilized in short-packet communications.

A. Short Packet Structure

In short-packet communications, a packet is simply composed of control information and payload. To facilitate the comparison between OB-based and conventional short-packet communications, all short packets are assumed to be of the same blocklength in this work. The short packet structure introduced by [1] is shown in Fig. 1, where the 'data' part at the physical (PHY) layer consists of the symbols coded and modulated by payload bits. The payload bits are produced by channel coding according to the DU from the media-access-control (MAC) layer and higher layers. The DU in a short packet can be expressed by a $1 \times K$ vector as

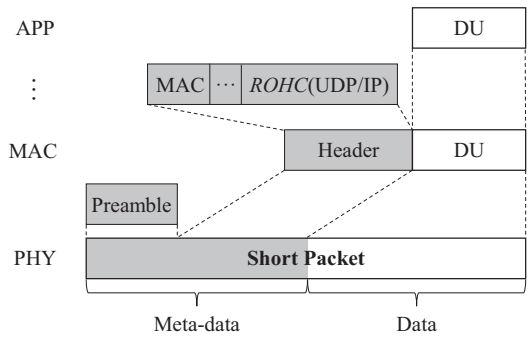


Fig. 1. The structure of a short packet.

$$\mathbf{v} = [v_1, v_2, \cdots, v_K], \tag{1}$$

where $v_k \in \{0, 1\}$ denotes the k^{th} information bit in the DU, $k = 1, 2, \cdots, K$, and K is the total number of bits in the DU.

The 'meta-data' part at the PHY layer consists of the preamble used for synchronisation and channel estimation as well as the symbols coded and modulated by control information bits. The control information bits are generated according to the header from the MAC layer and higher layers. The header can be expressed by a $1 \times L$ vector as

$$\mathbf{u} = [u_1, u_2, \cdots, u_L], \tag{2}$$

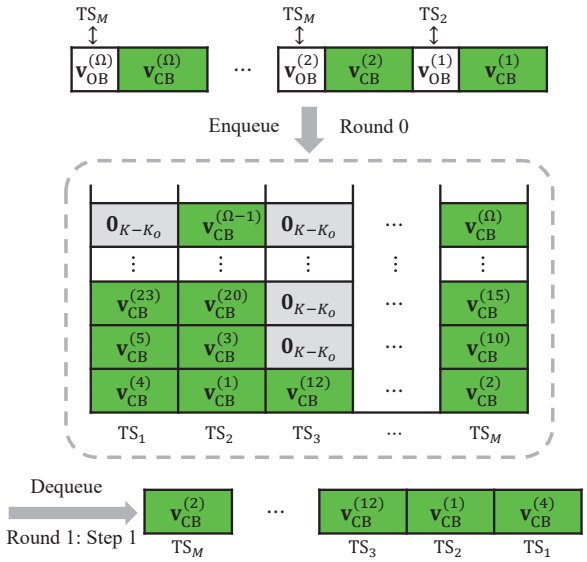
where $u_l \in \{0, 1\}$ is the l^{th} bit in the header, $l = 1, 2, \cdots, L$, and L is the total number of bits in the header.

For example, in the 5G New Radio, short packets are structured at transport layer by the User Datagram Protocol (UDP) that is typically used in the applications that require little overhead and accordingly achieve higher network throughput. Note that, the sequence number is an essential for IP-based communications whether TCP or UDP is adopted at the transport layer. Different from a TCP packet whose header contains a 32-bit sequence number added by the transport layer, the sequence number of a connectionless UDP packet is generated by the application (APP) layer to avoid packet loss. Also, a UDP/IP header can be compressed into at least two bytes by the Robust Header Compression (ROHC) to shorten overhead [28].

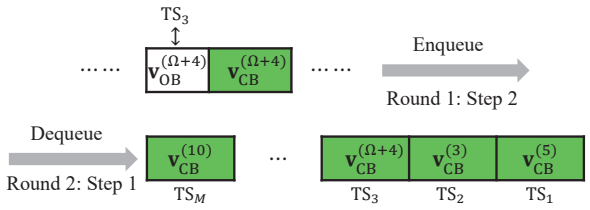
From the PHY layer view, a short packet delivered through wireless links is represented by a vector containing N symbols as

$$\mathbf{x} = \mathcal{C}\left([\mathbf{u}, \mathbf{v}]\right) = [x_1, x_2, \cdots, x_N], \tag{3}$$

where $\mathcal{C}(\cdot)$ denotes the function of PHY layer signal processing, including channel coding and modulation. The coding rate is $(L+K)/N$ within the blocklength N, to guarantee the same



(a) Storage Initialisation



(b) Enqueue and Dequeue

Fig. 2. An M-queue storage of CB vectors in OB-based packets.

reliability achieved by the meta-data and the payload data [1]. Since the header length L is comparable to the DU length K and cannot be omitted in short-packet communications, the payload rate K/N is used as a metric to gauge the effective delivery excluding the meta-data.

At the destination, received packets can be ordered according to their sequence numbers. Therefore, a transmitter does not need to deliver packets in the order of their sequence numbers, and extra information can be carried by the order of packet delivery. This is the foundation for the OB concept proposed in IP-based communications [24].

B. Opportunistic-Bits in Short-Packet Communications

The OB concept in [24] is proposed to improve the spectral efficiency for the delivery of massive IP-based packets from a transmitter to a receiver over a single link. Based on this concept, a portion of information bits in the DU, referred to as OBs, are used to indicate the TS index. The other information bits in the DU, referred to as CBs, are transmitted in the TS mapped by the OBs. That is, the transmitter delivers a bulk of packets in the order determined by their OBs.

To utilize this concept in short-packet communications, the DU of a short packet, given by (1), is divided into two parts as

$$\mathbf{v} = [\mathbf{v}_{\rm OB}, \mathbf{v}_{\rm CB}], \tag{4}$$

where the OB vector $\mathbf{v}_{\rm OB} = [v_1, v_2, \cdots, v_{K_o}]$ contains K_o OBs and the CB vector $\mathbf{v}_{\rm CB} = [v_{K_o+1}, v_{K_o+2}, \cdots, v_K]$ contains $K - K_o$ CBs.

Since there are K_o bits in the OB vector, the number of $\mathbf{v}_{\rm OB}$ variations is $M = 2^{K_o}$, which is the maximum number of TSs mapped by the OB vector. Therefore, we define M TSs as a group and the index of a TS in this group is $m \in \{1, 2, \cdots, M\}$. The OB vector $\mathbf{v}_{\rm OB}$ is mapped onto the index of a TS, m, and the mapping is a one-to-one correspondence, e.g., $m = 1 + \mathcal{D}(\mathbf{v}_{\rm OB})$.

Then, the CB vector $\mathbf{v}_{\rm CB}$ will be loaded in the $m^{\rm th}$ queue of a storage, $m = 1, 2, \cdots, M$. The storage, composed of M queues to cache DUs, is referred to as an M-queue storage. The loading process of CB vectors is shown in Fig. 2, where Ω CB vectors are to be loaded in the M-queue storage, $\Omega > M$, and the M queues pertain to the M TS indices mapped by K_o OBs. The superscript (ω) of $\mathbf{v}_{\rm CB}^{(\omega)}$ denotes the sequence number of the DU where the CB vector originates, $\omega = 1, 2, \cdots, \Omega$. Each storage unit is initialized by $K - K_o$ bits with value 0, i.e., the $1 \times (K - K_0)$ zero vector $\mathbf{0}_{K-K_o}$, and the initial bits hold there if none of the OB vectors pertaining to the enqueued CB vectors matches the TS index, i.e., the queue index. The TS bearing a CB vector is referred to as a *loaded TS*, and the TS bearing a zero vector is referred to as an *unloaded TS*.

We remark that, the physical space of the M-queue storage is only used for the Ω CB vectors. In practice, the beginning and the end of a queue are tracked by their addresses. That is, the zero vectors $\mathbf{0}_{K-K_o}$ shown in Fig. 2 do not exist in the physical space, and they only appear in the dequeue operation.

Upon the allocation of Ω CB vectors into the storage, the dequeue operation starts with the bottom row in the first round. From the first queue to the $M^{\rm th}$ queue, the bottom row of each one is a CB vector or a zero vector $\mathbf{0}_{K-K_o}$, and these M vectors are independent of each other. In a round, a number of CB vectors are dequeued and then the same number of upcoming CB vectors are enqueued at the top of this storage. Such dequeue and enqueue operations are repeated round by round to maintain the total number of CB vectors contained in the M-queue storage, Ω.

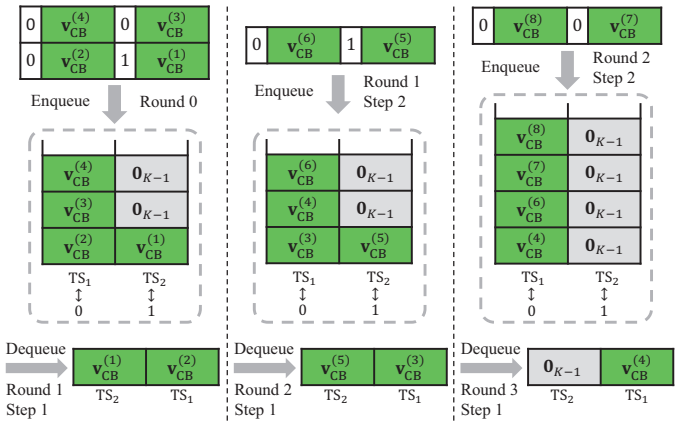
For example, the loading of $\Omega = 4$ CB vectors into a two-queue storage is illustrated in Fig. 3(a), where the OB size $K_o = 1$. If the OB in a packet $\mathbf{v}_{\rm OB} = [0]$, its CB $\mathbf{v}_{\rm CB}$ is loaded in TS 1. If $\mathbf{v}_{\rm OB} = [1]$, $\mathbf{v}_{\rm CB}$ is loaded in TS 2.

From the M-queue storage, the rescheduled CB vectors and the empty vectors $\mathbf{0}_{K-K_o}$ are packed with corresponding headers to generate short packets at the PHY layer. Especially, to deal with the empty vector $\mathbf{0}_{K-K_o}$, a pseudo header is packed with it to form a pseudo packet.

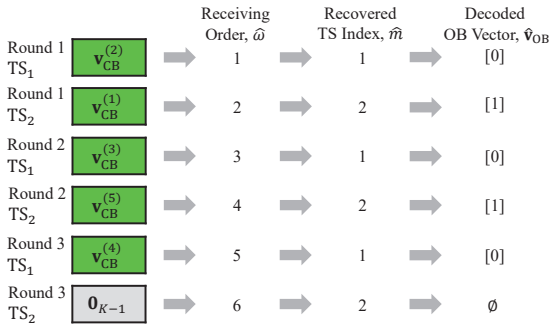
As the OB vector is removed from a DU and the DU is replaced by its CB vector, the short packet at the PHY layer is referred to as an OB-based packet and expressed as

$$\mathbf{x}_o = \mathcal{C}\left([\mathbf{u}, \mathbf{v}_{\rm CB}]\right) = [x_1, x_2, \cdots, x_{N_o}], \tag{5}$$

where N_o is the blocklength at the PHY layer for the physical transmission of $L + K - K_o$ bits in the OB-based packet $[\mathbf{u}, \mathbf{v}_{\rm CB}]$. Note that, OB vectors $\mathbf{v}_{\rm OB}$ will not be transmitted through physical channels, which leads to higher resource utilisation efficiency.



(a) Transmitter



(b) Receiver

Fig. 3. The OB-based design using a two-queue storage with $\Omega = 4$.

In short-packet communications, short channel codes are utilised in the channel coding and every packet is encoded independently [29], [30]. Therefore, the receiver can decode a packet immediately once receiving it. The information mapped by the CB vector in a packet, $\mathbf{v}_{\mathrm{CB}}$, is demodulated and decoded in the conventional way. The information conveyed by the OB vector in a packet, $\mathbf{v}_{\mathrm{OB}}$, is recovered through the index calculation of the TS when the packet is delivered, which is solely determined by the receiving order of the packet, $\hat{\omega}$. The TS index of the $\hat{\omega}^{\mathrm{th}}$ packet received at the receiver is calculated using $\hat{m} = ((\hat{\omega} - 1) \mod M) + 1$. Then, the information conveyed by the OB vector of this packet is obtained by $\hat{\mathbf{v}}_{\mathrm{OB}} = \mathcal{B}(\hat{m} - 1)$.

An illustration of the information recovery from the OB vectors is shown in Fig. 3(b), for receiving the packets transmitted from the two-queue storage given by Fig. 3(a).

In addition, a pseudo packet composed of the zero vector $\mathbf{0}_{K-K_o}$ and a pseudo header will be dropped by the receiver, because it cannot pass the header check. Eventually, the destination node reorders packets according to their sequence numbers.

III. Maximal Payload Rate

In this section, the maximal payload rate of OB-based short-packet communications is formulated in the finite blocklength regime. As introduced in Section II-A, the payload rate K/N is defined as the ratio of the DU size K to the blocklength N.

Since Shannon capacity deals with the transmissions over sufficiently large blocklength at arbitrarily small error probability, it is not suitable for short-packet communications. Herein, we utilize the analysis framework in the finite blocklength regime, specifically based on the normal approximation for nonasymptotic bounds of the maximum number of information bits that can be encoded [14, Eq. (296)]. Given the blocklength N, the block error probability ϵ and the signal-to-noise power ratio (SNR) ρ in AWGN channels, the normal approximation to the maximum number of information bits conveyed over the block is denoted by this normal approximation is denoted by

$$D(N, \epsilon, \rho) \approx NC(\rho) - \sqrt{NV(\rho)}Q^{-1}(\epsilon) + \frac{\log_2 N}{2}, \quad (6)$$

where the channel capacity $C(\rho) = (1/2)\log_2(1+\rho)$, the channel dispersion $V(\rho) = (\rho/2)[(\rho+2)/(\rho+1)^2](\log_2 e)^2$, and $Q^{-1}(\cdot)$ is the inverse function of $Q(x) = \int_x^\infty (1/\sqrt{2\pi})\exp(-t^2/2)dt$.

From [14, pp. 27], we notice that the normal approximation (6) can be fairly accurate when $D(N, \epsilon, \rho)/N > 0.8C(\rho)$. The main objective of this work is to qualitatively compare our OB-based transmissions with conventional transmissions in the same condition, for which we expect the normal approximation to be sufficiently accurate.

The maximal coding rate in the finite blocklength regime, $D(N, \epsilon, \rho)/N$, is lower than the channel capacity and can be derived using the information density $i(x;y) = \log_2 \frac{dP_{XY}}{d(P_X P_Y)}(x, y)$ [14, Eq. (3)], as the channel capacity $C(\rho)$ is the expectation of $i(x;y)$ and the channel dispersion $V(\rho)$, used to measure the channel's stochastic variability, is the variance of $i(x;y)$.

To transmit a DU of length K with a header of length L added based on the maximal coding rate, we have $L + K = D(N, \epsilon, \rho)$, where N is the minimum blocklength required to transmit $L + K$ bits at the target packet error probability ϵ and the SNR ρ over AWGN channels. Hence, the maximal payload rate of conventional short-packet communications is achieved at

$$R_c(L, N, \epsilon, \rho) = \frac{D(N, \epsilon, \rho) - L}{N}, \quad (7)$$

where the maximal payload length, i.e., the DU size, is given by

$$K = D(N, \epsilon, \rho) - L. \quad (8)$$

For OB-based short-packet communications, there are $K - K_o$ bits in the DU with an L-bit header added for the delivery of K-bit payload. Based on (5), we have $L + K - K_o = D(N_o, \epsilon, \rho)$ in the case of maximal coding rate, where N_o is the minimum blocklength required for the transmission of an L-bit header plus $K - K_o$ CBs. Apparently, $N_o < N$. Since $L+K = D(N, \epsilon, \rho)$ holds in conventional transmissions, we have $D(N, \epsilon, \rho) - K_o = D(N_o, \epsilon, \rho)$. Then, N_o can be expressed as

$$N_o(K_o, N, \epsilon, \rho) = D^{-1}(D(N, \epsilon, \rho) - K_o, \epsilon, \rho), \quad (9)$$

where $D^{-1}(k, \epsilon, \rho)$ is the inverse of the function $N \mapsto D(N, \epsilon, \rho)$ for fixed ϵ and ρ. Note that, $D(N, \epsilon, \rho)$ is given

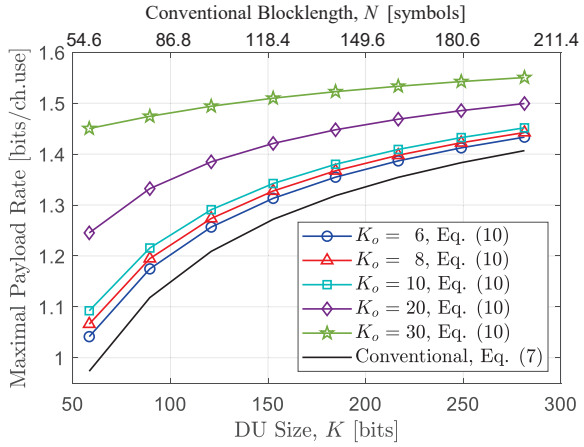


Fig. 4. Maximal payload rates of OB-based short-packets, (10), at the SNR $\rho = 10$dB, for $\epsilon = 10^{-3}$ and $L = 24$.

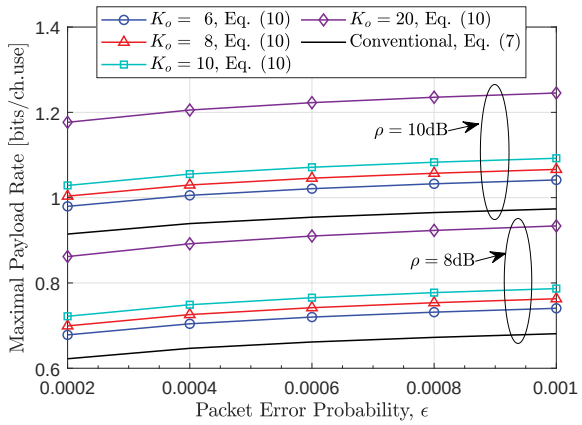


Fig. 5. Maximal payload rates of OB-based short-packets, (10), versus target packet error probability ϵ, at $\rho = 8$dB and 10dB, for $L = 24$ and $N = 60$.

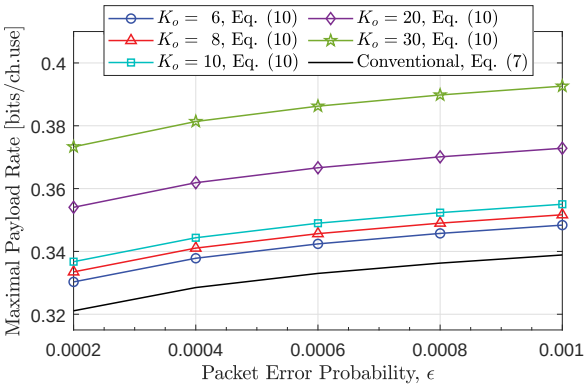


Fig. 6. Maximal payload rates of OB-based short-packets, (10), versus target packet error probability ϵ, at $\rho = 0$dB, for $L = 24$ and $N = 500$.

in (6), and $D^{-1}(k, \epsilon, \rho)$ denotes the minimum blocklength required for the delivery of k bits at the target packet error probability ϵ and the SNR ρ. Since $D(N, \epsilon, \rho)$ is a monotonic function, its inverse function exists. In our work, $D^{-1}(k, \epsilon, \rho)$ is obtained by searching paired inputs and outputs of (6) through numerical calculations.

Since the physical transmission of $K - K_o$ CBs in an OB-based packet conveys K payload bits, the payload rate of OB-based short-packet communications is K/N_o. Thus, based on (8) and (9), the maximal payload rate of OB-based short-packet communications is achieved at

$$R_o(K_o, L, N, \epsilon, \rho) = \frac{D(N, \epsilon, \rho) - L}{D^{-1}(D(N, \epsilon, \rho) - K_o, \epsilon, \rho)}, \quad (10)$$

given the number of OBs, K_o, and the target packet error probability ϵ at SNR ρ over AWGN channels.

In the following, the maximal payload rate of OB-based short-packet communications, given by (10), is compared with that of conventional short-packet communications, given by (7). To get the numerical results, the minimum blocklength N in the conventional design is set firstly, and then the maximal DU size K is obtained using N based on (8).

In Fig. 4, the maximal payload rates of OB-based short-packets given by (10) are compared with those of conventional short-packets given by (7), versus the number of information bits in a DU, K, at the SNR $\rho = 10$dB, for the header length $L = 24$ bits (i.e., 3 bytes) and the OB size $K_o = 6, 8, 10, 20, 30$. The target packet error probability ϵ is set to 10^{-3}. As shown in this figure, the maximal payload rate of OB-based transmissions is improved upon increasing the OB size and always higher than that of conventional transmissions. Note that, this improvement has to be limited by the header which takes about $1/8$ to $2/5$ of a short packet herein. If the OBs takes a larger proportion in a packet, increasing the OB size will lead to higher improvement of maximal payload rate.

In addition, the maximal payload rates of OB-based short-packets are plotted in Fig. 5, versus the packet error probability ϵ at the SNR $\rho = 8$dB and 10dB. For the sake of comparison, the maximal payload rates of conventional short-packets are also shown. Herein, the header length $L = 24$ and the conventional blocklength $N = 60$. This figure confirms general impacts of SNR and target packet error probability on the maximal payload rate. If we loosen the requirement on the target packet error probability or improve the SNR, higher maximal payload rates are achieved in both OB-based and conventional transmissions and, moreover, the payload rate gain obtained by OB-based packets over conventional ones increases. As anticipated based on the normal approximation $D(N, \epsilon, \rho)$ given by (6), the impact of the SNR ρ is greater than that of the packet error probability ϵ on the maximal payload rates.

At a low SNR, i.e., $\rho = 0$dB, the maximal payload rates are plotted versus the packet error probability ϵ in Fig. 6, where the blocklength N is set to 500 symbols, because the delivery of an L-bit header needs $\frac{L}{L+K} D^{-1}(L + K, \epsilon, \rho)$ symbols at the PHY layer and this number is quite large at low SNRs. As shown in this figure, the maximal payload rates of short-packet communications benefit from the OB-based design at a low SNR as well.

IV. Time-Slot Load Efficiency

As shown in Subsection II-B, an unloaded TS, i.e., the packet bearing $\mathbf{0}_{K-K_o}$, occurs because there is no OB vector mapped onto the TS index. Specifically for short-packet communications with the limitation on transceivers' buffer size,

the probability of no OB match for the TS index is a main concern on the feasibility of the OB paradigm.

In this section, we formulate the TS load efficiency, which is defined as the ratio of the number of loaded TSs to the total number of TSs involved in a loading process that is modelled by the M-queue storage in Fig. 2. In this M-queue storage, the random evolution of the number of CB vectors in each queue is a Repeated Balls-into-Bins (RBB) process, which is an ergodic finite state Markov chain with stationary distribution [31].

In OB-based communications, all the $M = 2^{K_o}$ possible variations of a K_o-bit OB vector $\mathbf{v}_{\mathrm{OB}}$ are assumed to occur at the same probability of $1/M$. Since a packet is transmitted in the TS mapped by its OB, the OB vector determines the queue where the CB vector is stored. Therefore, the probability that a CB vector $\mathbf{v}_{\mathrm{CB}}$ is stored into any of the M queues is $1/M$.

The RBB process of OB-based loading CB vectors is started by storing Ω CB vectors into M queues of the storage uniformly at random, $\Omega > M$. The storage initialisation is defined as Round 0, shown in Fig. 2(a). Subsequently, each round is composed of two steps as shown in Fig. 2(b).

- Step 1: The M items in the bottom row of the M-queue storage are dequeued in the order of the TS index of each queue. In detail, assuming there are λ empty queues in the storage, the $M-\lambda$ CB vectors and the λ zero vectors in the bottom row are dequeued. Afterwards, all items in the storage move one row down and, thus, the second bottom row before the dequeue operation becomes the bottom one.
- Step 2: $M-\lambda$ upcoming CB vectors are enqueued into the storage. Each of them is allocated to the queue mapped by its corresponding OB vector, which is a uniformly random allocation as the $M-\lambda$ OB vector variations are equiprobable.

As such, at the end of each dequeue-and-enqueue round, the total number of CB vectors contained in the storage is maintained at a constant Ω. The state of the RBB process for OB-based loading CB vectors at the end of Round r is denoted by $\mathbf{s}_r = [q_1, q_2, \cdots, q_M]$, where q_m represents the number of CB vectors in the m^{th} queue, $m = 1, 2, \cdots, M$. More specifically, $\sum_{m=1}^{M} q_m = \Omega$ always holds at the end of each round.

A. Stationary Distribution of Unloaded TSs

The number of unloaded TSs at the end of Round r is denoted by a random variable $\Lambda_r = \mathrm{num}(\mathbf{s}_r, 0)$, $r = 0, 1, 2, \cdots$, where $\mathrm{num}(\mathbf{x}, y)$ stands for the number of entries in the vector $\mathbf{x}$ that are equal to y. Thus, the number of empty queues in the storage at the end of Round r is $\Lambda_r \in \mathbb{M} = \{0, 1, \cdots, M-1\}$, where $\mathbb{M}$ is the sample space of Λ_r. Since the RBB process is not reversible and its explicit formula is unknown [32], we investigate the random variable Λ_r instead of the state $\mathbf{s}_r$ to work out the stationary distribution of Λ_r over states. An approximation of the probability transition matrix will be formulated for Λ_r to establish its stationary distribution.

Firstly, the initial distribution of Λ_r, i.e., Λ_0, is derived as follows. The sample space for allocating Ω CB vectors into M queues contains M^{Ω} outcomes. The number of event samples where no queue is empty upon the allocation of Ω CB vectors into M queues is given by the Stirling number of the second kind [33, pp. 824–825], expressed as

$$S(\Omega, M) = \frac{1}{M!} \sum_{j=1}^{M} (-1)^{M-j} \binom{M}{j} j^{\Omega}, \tag{11}$$

where j is an intermediate variable, denoting the number of non-empty queues.

Hence, the number of event samples where no queue is empty upon the allocation of Ω CB vectors into $M-\lambda$ queues can be denoted by $S(\Omega, M-\lambda)$. Since there are $M!/\lambda!$ permutations of $M-\lambda$ queues taken from M different queues, the number of event samples where λ queues are empty upon the allocation of Ω CB vectors into M queues is obtained by

$$U_0(\Omega, M, \lambda) = \frac{M!}{\lambda!} S(\Omega, M-\lambda). \tag{12}$$

For the initial distribution, the probability of λ empty queues in Round 0 is $\mathbb{P}[\Lambda_0 = \lambda] = U_0(\Omega, M, \lambda)/M^{\Omega}$, and $\sum_{\lambda=0}^{M-1} \mathbb{P}[\Lambda_0 = \lambda] = 1$.

The number of empty queues after the dequeue operation in Step 1 of Round r is denoted by Λ'_r. In Round 1, $M - \Lambda_0$ CB vectors are dequeued, so the number of empty queues after Step 1 in this round, $\Lambda'_1 \geqslant \Lambda_0$. Then, in Step 2 of this round, $M - \Lambda_0$ CB vectors are allocated into the M queues. Similar with the initial distribution, the probability that κ queues have null input after Step 2 of Round 1 is given by $U_0(M - \Lambda_0, M, \kappa)/M^{M-\Lambda_0}$. Obviously, the probability that none of the queues have null input in Round 1 is lower than the probability that no queue is empty in Round 0, as $U_0(M - \Lambda_0, M, 0)/M^{M-\Lambda_0} < \mathbb{P}[\Lambda_0 = 0]$. Therefore, the mean number of empty queues, $\mathbb{E}[\Lambda_r] = \sum_{\lambda=0}^{M-1} \mathbb{P}[\Lambda_r = \lambda]\lambda$, is monotonically increasing in r until this RBB process gets stationary. As such, $\mathbb{E}[\Lambda_0]$ in Round 0 pertains to the minimum $\mathbb{E}[\Lambda_r]$, although the initial distribution is independent of the stationary distribution.

Secondly, the probability transition matrix of Λ_r is analysed, which is defined as an $M \times M$ matrix $\mathbf{P}^{(r)} = (p_{\nu\lambda}^{(r)})_{M\times M}$, in Round r. More concretely, we have

$$\mathbf{P}^{(r)} = \begin{bmatrix} p_{00}^{(r)} & \cdots & p_{0,M-1}^{(r)} \\ \vdots & \ddots & \vdots \\ p_{M-1,0}^{(r)} & \cdots & p_{M-1,M-1}^{(r)} \end{bmatrix}, \tag{13}$$

where $p_{\nu\lambda}^{(r)} = \mathbb{P}[\Lambda_{r+1} = \lambda | \Lambda_r = \nu]$ is the $(\nu, \lambda)^{\mathrm{th}}$ entry of $\mathbf{P}^{(r)}$, $\nu, \lambda \in \mathbb{M}$, denoting the conditional probability of the empty-queue number transition from ν in Round r to λ in Round $r+1$. The calculation of the transition probability is given by

$$p_{\nu\lambda}^{(r)} = \sum_{\delta=\max\{\nu-\lambda,0\}}^{\min\{M-\lambda-1, M-\nu\}} \Big(\mathbb{P}[\Lambda'_{r+1} = \lambda + \delta | \Lambda_r = \nu] \times \mathbb{P}[\Lambda_{r+1} = \lambda | \Lambda'_{r+1} = \lambda + \delta, \Lambda_r = \nu] \Big), \tag{14}$$

where $\mathbb{P}[\Lambda'_{r+1} = \lambda + \delta | \Lambda_r = \nu]$ denotes the probability that the number of empty queues turns into $\lambda + \delta$ after the dequeue

operation in Step 1 of Round $r+1$ given $\Lambda_r = \nu$, and $\mathbb{P}[\Lambda_{r+1} = \lambda | \Lambda'_{r+1} = \lambda + \delta, \Lambda_r = \nu]$ denotes the probability that δ out of $\lambda + \delta$ empty queues have input in Step 2 of Round $r+1$ upon the allocation of $M - \nu$ CB vectors into M queues.

In detail, the probability

$$\mathbb{P}[\Lambda'_{r+1} = \lambda + \delta | \Lambda_r = \nu] = \mathbb{P}[\mathrm{num}(\mathbf{s}_r, 1) = \lambda - \nu + \delta | \Lambda_r = \nu], \tag{15}$$

where the item on the right-hand side represents the probability of the event that there are $\lambda - \nu + \delta$ 1's in $\mathbf{s}_r$ given that there are ν 0's in $\mathbf{s}_r$. This event implies that among all the $M - \nu$ non-empty queues at the end of Round r, each of $\lambda - \nu + \delta$ queues contains a single CB vector and each of the other $M - \lambda - \delta$ queues contains at least two CB vectors.

Hence, after the dequeue operation in Step 1 of Round $r+1$, $\lambda - \nu + \delta$ more queues get empty and the total number of empty queues turns into $\lambda + \delta$. The probability $\mathbb{P}[\mathrm{num}(\mathbf{s}_r, 1) = \lambda - \nu + \delta | \Lambda_r = \nu]$ in Round 0 is distinct owing to the random allocation of Ω CB vectors into M queues, and (15) can be derived as

$$\mathbb{P}[\Lambda'_1 = \lambda + \delta | \Lambda_0 = \nu] = \mu(\Omega, M - \nu, \lambda - \nu + \delta), \tag{16}$$

where

$$\mu(x, y, z) = \frac{\binom{x}{z}\binom{y}{z} z! U_1(x - z, y - z)}{U_0(x, y, 0)} \tag{17}$$

denotes the probability that allocating x CB vectors into y queues and among these queues, each of z queues contains a single CB vector while each of the other $y - z$ queues contains more than one CB vectors. Moreover, $U_1(x, y) = y! \sum_{j=0}^{y-1} (-1)^j \binom{x}{j} S(x - j, y - j)$ denotes the number of event samples where every queue contains more than one CB vectors upon the allocation of x CB vectors into y queues.

In Round $r \geqslant 1$, the probability $\mathbb{P}[\mathrm{num}(\mathbf{s}_r, 1) = \lambda - \nu + \delta | \Lambda_r = \nu]$ is related with the number of rounds, r, and depends on the state $\mathbf{s}_r$, i.e., the distribution of the number of CB vectors in each queue. Since the explicit solution of this RBB process is still an open problem, we cannot obtain the distribution of $\mathbf{s}_r$ for $r \geqslant 1$. From the simulation results of $\mathbb{P}[\mathrm{num}(\mathbf{s}_r, 1) = \lambda - \nu + \delta | \Lambda_r = \nu]$, we found this probability is very close to $\mu(\Omega', M - \nu, \lambda - \nu + \delta)$, where Ω' is an integer smaller than Ω and decreases until convergence as the round index r increases. That is, $\mathbb{P}[\mathrm{num}(\mathbf{s}_r, 1) = \lambda - \nu + \delta | \Lambda_r = \nu]$ for stationary states is very close to the probability that each of $\lambda - \nu + \delta$ queues contains a single CB vector and each of the other $M - \lambda - \delta$ queues contains more than one CB vectors upon the allocation of less than Ω CB vectors into $M - \nu$ queues. By fitting the simulation results of $\mathbb{P}[\mathrm{num}(\mathbf{s}_r, 1) = \lambda - \nu + \delta | \Lambda_r = \nu]$ in the case that the distribution of Λ_r is stationary, Ω' can be calculated by the fitting function $\mathcal{F}(\Omega, M) = M \lfloor -78.15(\Omega/M)^{-0.01653} + 78.94 \rfloor$, where $\lfloor \cdot \rfloor$ denotes the greatest integer function.

Therefore, for stationary states, i.e., $\mathbf{P}^{(\tilde{r})} = \mathbf{P}^{(\infty)}$ in Round $\tilde{r}$, the probability in (15) is superseded by an empirical formula as

$$\begin{aligned} &\mathbb{P}[\Lambda'_{\tilde{r}+1} = \lambda + \delta | \Lambda_{\tilde{r}} = \nu] \\ &\quad \approx \mu(\mathcal{F}(\Omega, M), M - \nu, \lambda - \nu + \delta). \end{aligned} \tag{18}$$

For the probability $\mathbb{P}[\Lambda_{r+1} = \lambda | \Lambda'_{r+1} = \lambda + \delta, \Lambda_r = \nu]$ on the right-hand side of (14), the number of event samples that δ out of $\lambda + \delta$ empty queues have input in Step 2 of Round $r+1$ is $\binom{\lambda+\delta}{\delta} \sum_{j=\delta}^{M-\max\{\nu,\lambda\}} \binom{M-\lambda-\delta}{j-\delta} j! S(M - \nu, j)$, i.e., the sum over the event samples that j queues, including the δ empty queues, have input in the allocation of $M - \nu$ CB vectors and each of the j queues has at least one of the $M - \nu$ CB vectors. As the total number of event samples in this enqueuing is $M^{M-\nu}$, we have

$$\begin{aligned} &\mathbb{P}[\Lambda_{r+1} = \lambda | \Lambda'_{r+1} = \lambda + \delta, \Lambda_r = \nu] \\ &\quad = \frac{\binom{\lambda+\delta}{\delta} \sum_{j=\delta}^{M-\max\{\nu,\lambda\}} \binom{M-\lambda-\delta}{j-\delta} j! S(M - \nu, j)}{M^{M-\nu}}. \end{aligned} \tag{19}$$

Therefore, the transition probability in stationary states, i.e., of $\mathbf{P}^{(\tilde{r})} = \mathbf{P}^{(\infty)}$, is obtained by substituting (18) and (19) into (14), as

$$\begin{aligned} p_{\nu\lambda}^{(\tilde{r})} \approx \sum_{\delta=\max\{\nu-\lambda,0\}}^{\min\{M-\lambda-1, M-\nu\}} \Bigg[&\mu\left(\mathcal{F}(\Omega, M), M - \nu, \lambda - \nu + \delta\right) \\ &\times \frac{\binom{\lambda+\delta}{\delta} \sum_{j=\delta}^{M-\max\{\nu,\lambda\}} \binom{M-\lambda-\delta}{j-\delta} j! S(M - \nu, j)}{M^{M-\nu}} \Bigg], \end{aligned} \tag{20}$$

where the probability $\mu(x, y, z)$ is given by (17).

Finally, the stationary distribution of $\Lambda_{\tilde{r}}$ is denoted by a $1 \times M$ vector $\mathbf{p} = [p_0, p_1, \cdots, p_{M-1}]$, where the λ^{th} entry $p_\lambda = \mathbb{P}[\Lambda_{\tilde{r}} = \lambda]$, $\lambda \in \mathbb{M}$. This stationary distribution can be achieved by solving the equation set $\mathbf{p}\mathbf{P}^{(\tilde{r})} = \mathbf{p}$ and $\sum_{\lambda=0}^{M-1} p_\lambda = 1$.

Herein, we present a practical method to solve the stationary distribution. As stated by the Perron-Frobenius Theorem [34], if a Markov chain is irreducible and aperiodic, there is a unique stationary distribution $\mathbf{p}$. Moreover, the stationary transition matrix converges to a rank-one matrix and all rows in it are identical. Specifically, each row in the stationary transition matrix is the stationary distribution, which implies that the transition probabilities of moving from all states to any state are the same. Assuming the stationary distribution is achieved through α transitions, the α-step transition matrix is formed as an $M \times M$ matrix $\hat{\mathbf{P}} = (\mathbf{P}^{(\tilde{r})})^\alpha$, where α is a positive integer. To satisfy the stationary condition associated with the matrix formulation, all rows in $\hat{\mathbf{P}}$ are identical and equal to $\mathbf{p}$. In our experiments, α is set to 50, which guarantees that all the rows in $\hat{\mathbf{P}}$ are the same, for all values of K_o and Ω in the calculations. In this way, an accurate stationary distribution $\mathbf{p}$ is obtained, with the stationary probability

$$p_\lambda = \hat{\mathbf{P}}[\cdot, \lambda], \qquad \forall \lambda \in \mathbb{M}, \tag{21}$$

where $\hat{\mathbf{P}}[\cdot, \lambda]$ is the λ^{th} entry in an arbitrary row of $\hat{\mathbf{P}}$.

B. Impact of Unloaded TSs

In Fig. 7, the dequeue output examples of the M-queue storage in stationary states are presented to illustrate the delivery of CB vectors threaded with few unloaded TSs. The mean number of unloaded TSs in a dequeue operation is equal

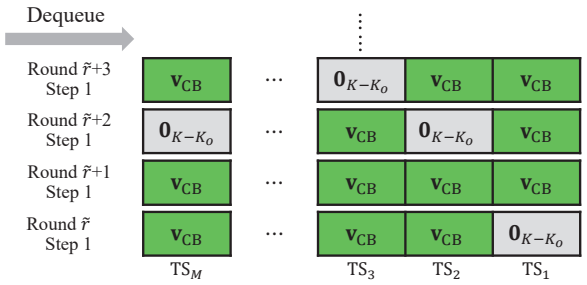


Fig. 7. Dequeue output examples of the M-queue storage in stationary states.

to the mean number of empty queues at the end of a round, and the TS load efficiency is the rate of loaded TSs.

The mean number of empty queues increases round by round until convergence when the RBB process gets stationary. Thus, we use the case in stationary states to evaluate the TS load efficiency of OB-based short-packet communications, where the mean number of unloaded TSs achieves its maximum at

$$\mathbb{E}[\Lambda_{\tilde{r}}] = \sum\nolimits_{\lambda=0}^{M-1} p_\lambda \lambda, \tag{22}$$

and the proportion of unloaded TSs in the stationary distribution is $\mathbb{E}[\Lambda_{\tilde{r}}]/M$.

As a result, the minimum of the TS load efficiency, i.e., in stationary states, is obtained by

$$\eta = 1 - \frac{\mathbb{E}[\Lambda_{\tilde{r}}]}{M} = 1 - 2^{-K_o} \sum_{\lambda=0}^{2^{K_o}-1} p_\lambda \lambda. \tag{23}$$

In the calculation of η, the transition probability $p_{\nu\lambda}^{(\tilde{r})}$ given in (20) is used for M^2 times to obtain the probability transition matrix $\mathbf{P}^{(\tilde{r})}$. Moreover, the complexity in the calculation of $p_{\nu\lambda}^{(\tilde{r})}$ is $O(M^3)$, as there are two layers of summation shown in (20) and another layer of summation in $S(\cdot,\cdot)$ is given by (11). Therefore, the complexity in the calculation of η is $M^2 \cdot O(M^3) = O(M^5) = O(2^{5K_o})$.

To verify the above derivations, we will compare them with simulation results on the stationary distribution of the unloaded TS number and the TS load efficiency of OB-based short-packet communications. Given M and Ω, 5000 simulations are run to exhibit the distribution of Λ_r from Round 0 to Round 10^4, which always converges within 10^4 rounds for $K_o \leqslant 10$. In Figs. 8, 9, and 10, the numerical results of (21), (22) and (23) are compared with their simulation results for $K_o = 4, 5, 6$. As K_o increases, the comparisons on η are presented in Fig. 11.

Fig. 8 depicts the stationary distribution of the events that the numbers of unloaded TSs are 0 and 1, versus the total packet number in the storage containing $M = 16, 32, 64$ queues, i.e., the number of OBs, $K_o = 4, 5, 6$, where the theoretical results are calculated using (21) based on the transition matrix approximation in (20). As shown in this figure, the theoretical and simulation results match well. Moreover, with the increase in Ω, the probability of $\lambda = 0$ converges to 1 while that of $\lambda = 1$ converges to 0. In other words, given M, the number of unloaded TSs is dramatically reduced upon

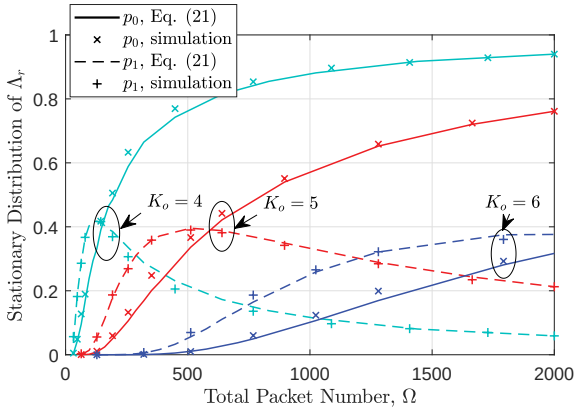


Fig. 8. Stationary distribution of $\Lambda_{\tilde{r}} = \lambda$, for the events $\lambda = 0$ and $\lambda = 1$, versus the total number of packets, with $K_o = 4, 5, 6$.

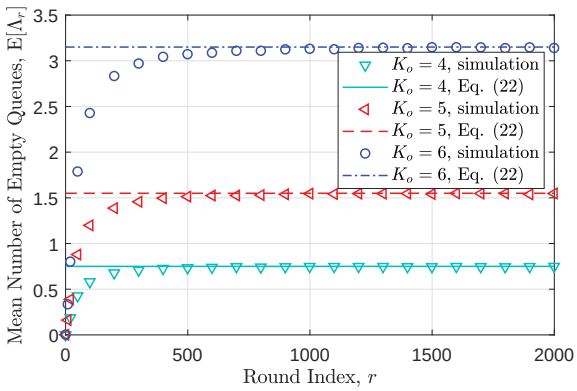


Fig. 9. The mean number of empty queues versus the round index r, for $\Omega/2^{K_o} = 10$ and $K_o = 4, 5, 6$.

increasing Ω, and there is no unloaded TS in massive access of short packets, i.e., in the case of large Ω. The main reason behind this is that a large number of packets, Ω, in the M-queue storage leads to sufficient rounds of loading CB vectors and stationary states, which guarantees a negligible number of unloaded TSs in the dequeue operation of each round.

Fig. 9 shows the evolution of $\mathbb{E}[\Lambda_r]$ versus the round index r, where $\mathbb{E}[\Lambda_r]$ increases along with the growth of the round number until meeting the convergence. From this figure, we may find that $\mathbb{E}[\Lambda_0]$ is the minimum of $\mathbb{E}[\Lambda_r]$, as justified by the initial distribution in Section IV-A, and the stationary case $\mathbb{E}[\Lambda_{\tilde{r}}]$ given by (22) is the upper bound on $\mathbb{E}[\Lambda_r]$. Hence, the TS load efficiency in stationary states of this RBB process, i.e., η given in (23), is the lower bound of the instantaneous TS load efficiency and we will use the lower bound η as the TS load efficiency in the following study. That is, we will consider the worst case.

Further, the theoretical and simulation results on the TS load efficiency are compared for the OB size $K_o = 4, 5, 6$ in Fig. 10, where the theoretical results are obtained from (23) and they are very close to the simulation results. This figure reveals that the TS load efficiency converges to 1 as the total packet number Ω increases. For the case of larger K_o, the TS load efficiency approaches 1 as Ω increases.

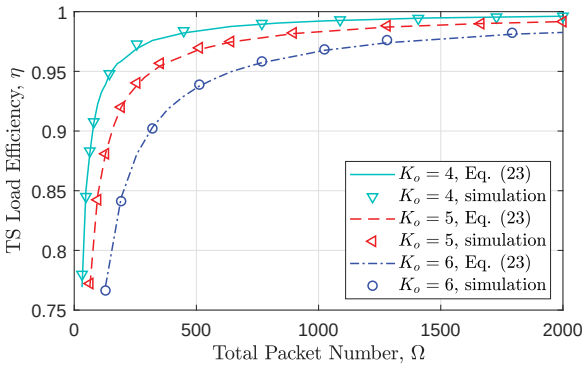


Fig. 10. TS load efficiency versus the total number of packets, for $K_o = 4, 5, 6$.

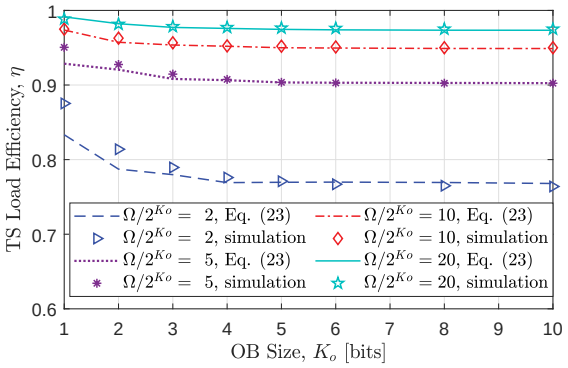


Fig. 11. TS load efficiency versus the OB size K_o, for $\Omega/2^{K_o} = 2, 5, 10, 20$.

In detail, the TS load efficiency is plotted as a function of K_o in Fig. 11, for the cases of $\Omega/2^{K_o} = 2, 5, 10, 20$, where the ratio $\Omega/2^{K_o}$ stands for the mean queue length in the M-queue storage. Since the complexity in the calculation of η given in (23) is $O(2^{5K_o})$, the results for $K_o > 10$ cannot be obtained. As is shown in this figure, the theoretical and simulation results of the TS load efficiency η match very well, except for the case of a low ratio $\Omega/2^{K_o}$ with a very small OB size, e.g., $\Omega/2^{K_o} = 2, 5$ with $K_o = 1, 2$. In the cases of $K_o \leqslant 5$, η decreases as K_o increases, for a given ratio $\Omega/2^{K_o}$. However, for K_o ranging from 6 to 10, η almost keeps the same value given a ratio $\Omega/2^{K_o}$. Based on this tendency, we predict that the TS load efficiency for $10 < K_o \leqslant 20$ is supposed to be the same as that at $K_o = 10$, given a ratio $\Omega/2^{K_o}$. Moreover, η is improved as the ratio $\Omega/2^{K_o}$ increases. However, setting a high ratio $\Omega/2^{K_o}$ is not economical concerning the storage cost. For example, the storage space for $\Omega/2^{K_o} = 20$ is twice that for $\Omega/2^{K_o} = 10$, while η is only improved from 0.95 to 0.975 in the case of $K_o = 10$.

V. Performance Analysis

In this section, the resource utilisation efficiency and performance of OB-based short-packet communications are investigated in the metrics of energy gain, goodput, and latency, based on the formulations of the maximal payload rate in Section III and the TS load efficiency in Section IV. With the aid of the performance analysis outcomes, how to select the OB size K_o in an efficient OB-based design is discussed.

Herein, the header length L of a packet is fixed to 24 bits, and the variation in blocklength N results from various DU sizes K.

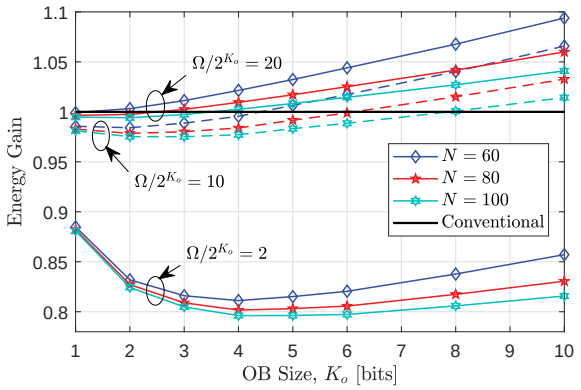


Fig. 12. Energy gain of OB-based short-packet communications, (24), versus the OB size K_o, at the SNR $\rho = 10$dB, for $\epsilon = 10^{-3}$, $L = 24$ and $\Omega/2^{K_o} = 2, 10, 20$.

A. Energy Gain

Without loss of generality, the average energy used to transmit a short packet conveying a K-bit DU is assumed to be a constant denoted by E. Thus, the symbol energy in a conventional short-packet of length N is E/N, and the symbol energy in an OB-based short-packet of length N_o is $\eta E/N_o$, where η is the TS load efficiency given in (23) and N_o is calculated using (9).

Given the same packet energy, the energy allocated to each symbol in an OB-based packet is likely more than that in a conventional one, because the OB-based packet length N_o is shorter than the conventional N, while a single DU is delivered through $1/\eta$ OB-based packets on average. Consequently, the energy gain achieved by OB-based communications over conventional communications, denoted by ξ_E, is defined as the ratio of OB-based symbol energy $\eta E/N_o$ to the conventional symbol energy E/N, expressed in detail by

$$\xi_E = \frac{\eta E/N_o}{E/N} = \frac{\eta N}{D^{-1}\left(D(N, \epsilon, \rho) - K_o, \epsilon, \rho\right)}. \tag{24}$$

We remark that, a higher energy gain means a smaller ratio N_o/η compared with N.

In Fig. 12, the energy gain of OB-based short-packet communications is plotted as a function of the OB size K_o, for the SNR $\rho = 10$dB and the header length $L = 24$, where the conventional blocklength $N = 60, 80, 100$, and the ratio $\Omega/2^{K_o} = 2, 10, 20$. This figure reveals that the energy gain of OB-based short-packets decreases upon increasing the blocklength of a packet. Besides, the energy gain is lower than 1 in the case of $\Omega/2^{K_o} = 2$ and higher than 1 in the case of $\Omega/2^{K_o} = 20$. In the case of $\Omega/2^{K_o} = 10$, the energy gain is higher than 1 if the OB size $K_o \geqslant 5$ in a shorter packet, e.g., $N = 60$, i.e., where the OB takes a larger proportion in a packet. The main reason behind this is that the energy gain

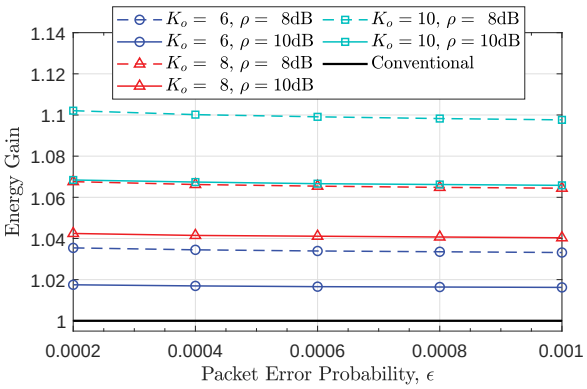


Fig. 13. Energy gain of OB-based short-packet communications, (24), versus target packet error probability ϵ, at the SNR $\rho = 8$dB and 10dB, for $L = 24$, $N = 60$, $K_o = 6, 8, 10$ and $\Omega/2^{K_o} = 10$.

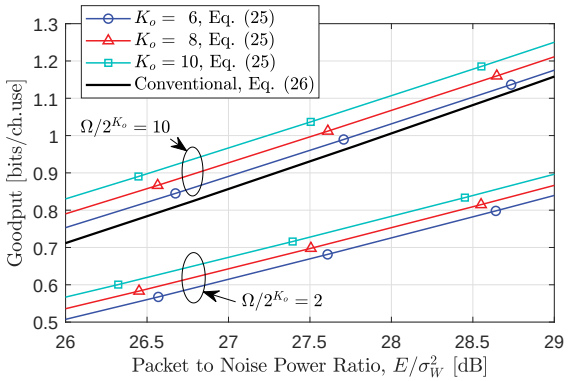


Fig. 14. Goodput comparisons between OB-based and conventional short-packet communications for $\epsilon = 10^{-3}$, $L = 24$, $N = 60$, $K_o = 6, 8, 10$ and $\Omega/2^{K_o} = 2, 10$.

ξ_E converges to the TS load efficiency η with the increase in blocklength N, which is implied by (24). That is, given K_o and Ω, the ratio N/N_o tends to 1 as N goes to infinity. Additionally, (24) implies that the energy gain ξ_E approaches N/N_o when η approaches 1. As shown in Fig. 11, $\eta = 0.975$ when $K_o = 10$ and $\Omega = 20 \times 2^{10}$. In this case, $\xi_E \approx N/N_o$ as shown in Fig. 12, where $\xi_E = 1.09, 1.06, 1.04$ given that $N = 60, 80, 100$ with $N_o = 53.5, 73.6, 93.7$.

Furthermore, Fig. 13 investigates the impacts of target packet error probability ϵ and SNR ρ on the energy gain of OB-based short-packet communications, ξ_E in (24), where the SNR ρ is set to 8dB and 10dB, for the OB size $K_o = 6, 8, 10$, the header length $L = 24$, and the conventional blocklength $N = 60$. The total number of short packets contained in the M-queue storage, $\Omega = 10 \times 2^{K_o}$, which is set based on the aforementioned analysis to achieve a good balance between the performance improvement and the storage cost. An interesting phenomenon in this figure is that the energy gain of OB-based short-packets increases upon reducing the SNR or meeting a more stringent requirement on the target packet error probability. This is because the maximum number of information bits transmitted by a short packet of length N decreases with the decline in ρ or ϵ and, thus, the OB-based blocklength N_o decreases as well. Consequently, the ratio $N/N_o(K_o, N, \epsilon, \rho)$ increases as ρ or ϵ decreases.

B. Goodput

To measure effective delivery of application-layer data in OB-based short-packet communications, the metric of goodput, i.e., the application-layer throughput excluding the metadata arising from control information, is investigated herein. Concerning the TS load efficiency, the maximal payload rate and the energy gain discussed above, the goodput of OB-based short-packet communications, denoted by G_o in the unit of [bits/channel use], is given by

$$G_o = (1-\epsilon)\eta R_o(K_o, L, N, \epsilon, \rho\xi_E), \tag{25}$$

where η, ξ_E and $R_o(K_o, L, N, \epsilon, \rho\xi_E)$ are the TS load efficiency in (23), the energy gain in (24) and the maximal coding rate in (10), respectively. Moreover, $(1-\epsilon)$ is the rate of correctly received packet, and η stands for the rate of loaded TSs.

For the purpose of comparison, the goodput of conventional short-packet communications, denoted by G_c in the unit of [bits/channel use], is expressed as

$$G_c = (1-\epsilon)R_c(L, N, \epsilon, \rho), \tag{26}$$

where $R_c(L, N, \epsilon, \rho)$ is given by (7).

Subsequently, the OB-based goodput (25) is compared with the conventional goodput (26). To begin with, (25) and (26) are plotted as functions of packet-to-noise power ratio E/σ_W^2 in Fig. 14, where σ_W^2 is the AWGN variance, i.e., equal to the power spectral density of the AWGN in the communication channel. Slightly elaborating further, the SNR $\rho = E_s/\sigma_W^2$, where E_s denotes the symbol energy. Therefore, we have $E/\sigma_W^2 = N\rho$ in conventional short-packet communications. As the length of an equivalent conventional packet is set to $N = 60$, we have E/σ_W^2 in [dB] $= \rho$ in [dB] $+$ 19dB. Besides, the target packet error probability $\epsilon = 10^{-3}$, the header length $L = 24$, the OB size $K_o = 6, 8, 10$, and the ratio $\Omega/2^{K_o} = 2, 10$. This figure reveals that the OB-based goodput is improved upon increasing the SNR ρ or the OB size K_o. The feature of goodput in this figure is a counterpart of energy gain in Fig. 12. When $\Omega/2^{K_o} = 10$, our OB-based design achieves higher goodput than the conventional. However, when $\Omega/2^{K_o} = 2$, the OB-based goodput is lower than the conventional goodput, mainly because the blocklength gain brought by the OB concept is used to compensate the low TS load efficiency in this case.

Then, the impacts of higher ratio $\Omega/2^{K_o}$ and larger blocklength on the OB-based goodput is investigated in Fig. 15, where the ratio $\Omega/2^{K_o} = 20$ and the conventional blocklength $N = 60, 120$. The other parameters are the same as those in Fig. 14. As shown herein, the OB-based goodput G_o is higher than the conventional G_c, when $\Omega/2^{K_o} = 20$. Moreover, G_o, G_c and the goodput increment $G_o - G_c$ decrease with the increase in N. The main reason behind this is that the TS load efficiency approaches 1, i.e., $\eta = 0.975$, in the case of $\Omega/2^{K_o} = 20$. As such, the OB-based goodput (25) is determined by the maximal payload rate R_o. As shown in Fig. 12, the energy gain ξ_E decreases upon increasing N,

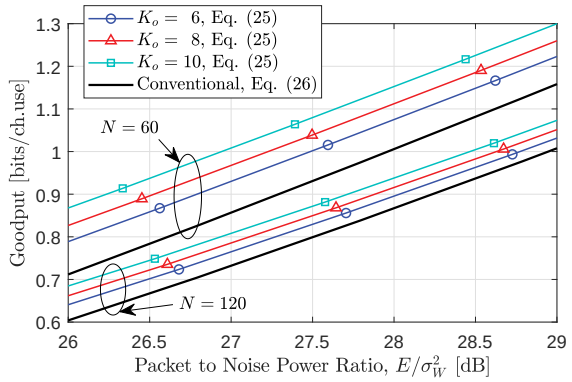


Fig. 15. Goodput comparisons between OB-based and conventional short-packet communications, for $\epsilon = 10^{-3}$, $L = 24$, $N = 60, 120$, $K_o = 6, 8, 10$ and $\Omega/2^{K_o} = 20$.

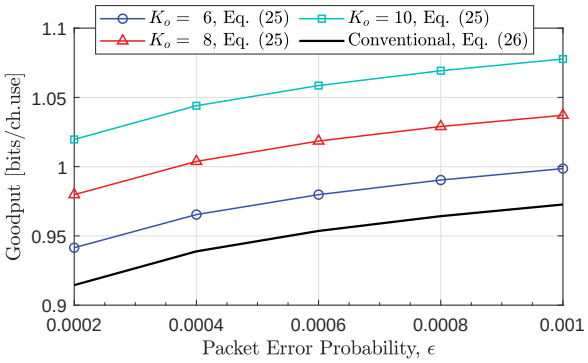


Fig. 16. Goodput comparisons between OB-based and conventional short-packet communications, for $\rho = 10$dB, $L = 24$, $N = 60$, $K_o = 6, 8, 10$ and $\Omega/2^{K_o} = 10$.

which leads to the reduction in R_o and $R_o - R_c$. Besides, the conventional goodput (26) is determined by R_c. The conventional symbol energy E/N decreases as N increases, which leads to the reduction in R_c as well. Consequently, the goodput gain achieved by OB-based design over the conventional relies on the tradeoff among N, ρ, and ξ_E.

Besides, from Fig. 14, we may notice that the goodput in the case of $\Omega/2^{K_o} = 10$ is much higher than that in the case of $\Omega/2^{K_o} = 2$. However, by comparing the case of $N = 60$ in Figs. 14 and 15, we may find that, increasing the ratio $\Omega/2^{K_o}$ from 10 to 20 only results in a small goodput increment. As the saturation emerges in the improvement of goodput, the ratio $\Omega/2^{K_o} = 10$ is a good setting for the OB-based system design, which can balance the resource utilisation efficiency and the number of storage units in the M-queue storage.

Furthermore, (25) and (26) are plotted as functions of the target packet error probability ϵ in Fig. 16, at the SNR $\rho = 10$dB, for the header length $L = 24$, the conventional blocklength $N = 60$, the OB size $K_o = 6, 8, 10$, and the ratio $\Omega/2^{K_o} = 10$. As shown in this figure, both the OB-based goodput and the conventional goodput increase when the requirement on target packet error probability gets looser, i.e., when ϵ gets larger. Moreover, to achieve the same goodput, OB-based short-packets can meet much more stringent requirement on the target packet error probability than conventional ones.

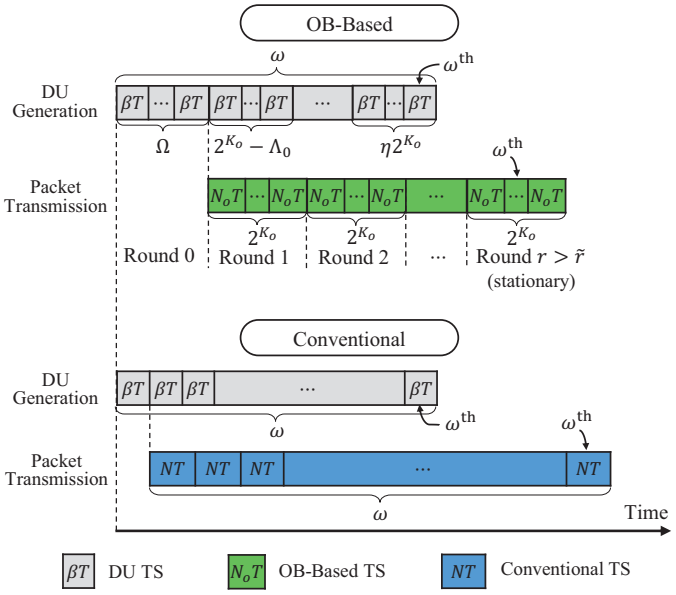


Fig. 17. DU generation and packet transmission in OB-based and conventional short-packet communications.

C. Latency

To investigate the impact of the OB-based design on the tradeoff between resource efficiency and latency, three types of latency in the OB-based short-packet communications are taken into account for a packet: *(i)* the average latency of a packet loaded in the longest queue, *(ii)* the average latency of a packet loaded in an empty queue, and *(iii)* the mean latency of a packet.

Without loss of generality, the OB-based and conventional communications are assumed to spend the same time in signal processing and channel coding at the PHY layer as well as in reading from and writing to the storage. Hence, the difference between their latencies is determined by the time in DU generation and packet transmission, which is shown in Fig. 17. Herein, data generation is assumed to be faster than the transmission, which agrees with the reality of short-packet communications. The generation TS duration of a K-bit DU is denoted by βT, where T is a unit time. As introduced in Section II, to deliver the same payload, i.e., a K-bit DU, the TS durations of an OB-based packet and a conventional packet are N_oT and NT, respectively, where $\beta < N_o < N$.

For a fair comparison between the OB-based and conventional short-packet communications, the latency of a packet is defined as the time elapsed from the generation to the delivery of its DU. As shown in Fig. 17, the OB-based initialisation needs to generate a bulk of Ω DUs and, thus, its initialisation time is $\Omega\beta T$. The conventional initialisation only needs to generate a single DU, because the transmission is started once the first DU is generated. As data generation is faster than the transmission, i.e., $\beta < N$, the transmission of the ω^{th} DU will not be started until the previous DUs has been transmitted.

Therefore, in the conventional transmission, the initialisation time is βT, and the latency of the ω^{th} DU is calculated using

$$\tau_{\omega}^{\text{C}} = \beta T + \omega N T - \omega \beta T = [\beta + \omega\,(N - \beta)]\,T. \tag{27}$$

Apparently, if $\omega < \Omega$, the latency of the ω^{th} DU in the OB-based transmission is likely longer than that in the conventional transmission, since the OB-based initialisation time $\Omega\beta T$ is likely longer than τ_{ω}^{C}.

In the following, we contrast the OB-based latency against the conventional τ_{ω}^{C} for the practical case of $\omega \gg \Omega$. Note that, compared with conventional transmissions, the main advantage of our OB-based design is the reduced blocklength, which leads to a shorter TS of each packet, i.e., $N_o T < NT$. However, the disadvantage of the OB-based design lies in the transmission order of packets. More specifically, the OB-based packets are not transmitted in the order labelled by their sequence numbers, $\omega = 1, 2, \cdots, \Omega$. As shown in Fig. 2, the transmission order of the m^{th} TS in Round r is numbered by $(r-1)M + m$, where $M = 2^{K_o}$ is the number of queues in the storage and $m = 1, 2, \cdots, M$. Some packets are transmitted earlier than their sequence numbers, e.g., $\mathbf{v}_{\text{CB}}^{(4)}$, $\mathbf{v}_{\text{CB}}^{(12)}$, $\mathbf{v}_{\text{CB}}^{(\Omega+4)}$, and some packets are transmitted later than their sequence numbers, e.g., $\mathbf{v}_{\text{CB}}^{(1)}$, $\mathbf{v}_{\text{CB}}^{(2)}$, $\mathbf{v}_{\text{CB}}^{(3)}$. However, the conventional packets are transmitted in the order labelled by their sequence numbers. The latency of the ω^{th} conventional packet is $\tau_{\omega}^{\text{C}} = \beta T + \omega N T - \omega \beta T$, as given in (27). In comparison to the conventional transmission, the OB-based design offers shorter latency for those packets transmitted earlier than their sequence numbers but longer latency for those transmitted later than their sequence numbers.

In the M-queue storage, after the dequeue step of a given round, some queues contain more CB vectors and some contain less. Then, in the enqueue step of this round, the packet loaded into the queue containing the most CB vectors pertains the longest possible latency and the one loaded into the queue containing the least CB vectors pertains to the shortest possible latency.

When the RBB process is stationary, the queue containing most CB vectors after the dequeue operation in Step 1 of a round is denoted by Queue $\tilde{m}$, and the number of CB vectors contained in this queue is $\max\{\mathbf{s}_{\tilde{r}}\}$, which is the length of the longest queue, in the M-queue storage. Then, the waiting time of the packet loaded into Queue $\tilde{m}$ in Step 2 of this round is $(\max\{\mathbf{s}_{\tilde{r}}\} - 1)M + \tilde{m}$, which is the longest possible latency. The mean length of the longest queue is denoted by $Q = \mathbb{E}\,[\max\{\mathbf{s}_{\tilde{r}}\}]$. If there are $\Omega = M$ packets loaded in the M-queue storage, we have $Q = O(\log_2 M)$ based on [31, Theorem 1]. If $\Omega > M$, the process of loading Ω CB vectors into the M-queue storage can be regarded as Ω/M RBB processes of loading M CB vectors into the storage. The worst case is that all of these Ω/M RBB processes have the maximum number of CB vectors loaded into the same queue. In this case, the longest queue contains, on average, $\mathbb{E}[\max\{\mathbf{s}_{\tilde{r}}\}] \approx (\Omega/M)O(\log_2 M)$ CB vectors. Hence, based on the form of $Q \approx (\Omega/2^{K_o})O(K_o)$, we express the mean length Q as a function of K_o by fitting simulation results over 5000 runs of the 2^{K_o}-queue storage. There are 20000 dequeue-and-enqueue rounds in each run to get the RBB process stationary. As a result, Q can be expressed as

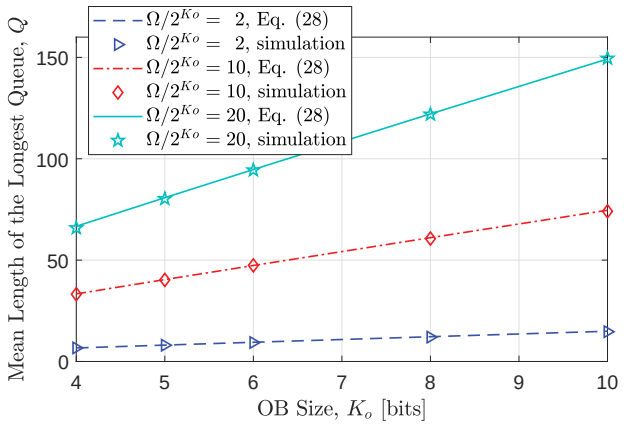


Fig. 18. Length of the longest queue in the 2^{K_o}-queue storage versus the OB size K_o, for $\Omega/2^{K_o} = 2, 10, 20$.

$$Q \approx \frac{\Omega}{2^{K_o}} \left(\frac{0.475}{K_o^{0.575}} + 0.62 \right) K_o. \tag{28}$$

The simulation results and the approximation (28) are compared in Fig. 18, which validates the reliability and accuracy of the approximation for various OB size K_o and various ratio $\Omega/2^{K_o}$.

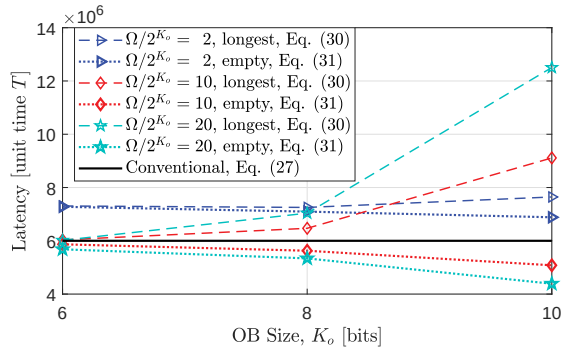
As aforementioned, the initialisation time in the OB-based transmission is $\Omega\beta T$. Moreover, as shown in Fig. 17, for the ω^{th} DU loaded in Round r, the transmission time of its previous $\omega - \Omega$ DUs is upper bounded by $(\omega - \Omega)N_o T/\eta$, because the actual TS load efficiency is higher than η before the loading process in the M-queue storage gets stationary. Thus, the waiting time of the ω^{th} packet, from the generation to the loading of its DU in the M-queue storage, is upper bounded by

$$\begin{aligned} \tau_{\omega}^{\text{wait}} &\leqslant (\Omega - \omega)\,\beta T + (\omega - \Omega)\,N_o T/\eta \\ &= [(\omega - \Omega)\,(N_o/\eta - \beta)]\,T. \end{aligned} \tag{29}$$

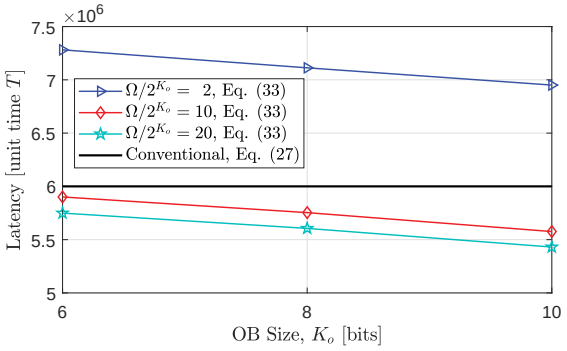
When the ω^{th} DU is loaded into the M-queue storage, the probability that its CB vector is loaded on the top of the longest queue is $1/M$. In this case, the DU will wait for $QM - M/2$ TSs, on average, to be dequeued. Hence, the average latency of the ω^{th} packet dequeued from the longest queue, denoted by $\tau_{\omega}^{\text{longest}}$, is upper bounded by

$$\begin{aligned} \tau_{\omega}^{\text{longest}} &= \tau_{\omega}^{\text{wait}} + (QM - M/2)\,N_o T \\ &\leqslant (\Omega - \omega)\,\beta T + \left[(\omega - \Omega)/\eta + (Q - 1/2)\,2^{K_o}\right] \\ &\quad \times D^{-1}\,(D(N, \epsilon, \rho) - K_o, \epsilon, \rho)\,T. \end{aligned} \tag{30}$$

Moreover, the probability that the ω^{th} DU's CB vector is loaded at the bottom of an empty queue is $\mathbb{E}[\Lambda_{\tilde{r}}]/M = 1 - \eta$. In this case, the DU will wait for $M/2$ TSs, on average, to be dequeued. Hence, the average latency of the ω^{th} packet



(a)



(b)

Fig. 19. Latency of the 10^5th DU in OB-based short-packet communications versus OB size, excluding initialisation time, for $\rho = 10$dB, $\epsilon = 10^{-3}$, $L = 24$, $N = 60$. (a) Average Latencies of the packet dequeued from the longest and the empty queues. (b) Mean latency.

dequeued from an empty queue, denoted by $\tau_\omega^{\text{empty}}$, is upper bounded by

$$\begin{aligned}\tau_\omega^{\text{empty}} &= \tau_\omega^{\text{wait}} + MN_oT/2 \\ &\leqslant (\Omega-\omega)\,\beta T + \left[(\omega-\Omega)/\eta + 2^{K_o-1}\right] \\ &\quad \times D^{-1}\left(D(N,\epsilon,\rho) - K_o, \epsilon, \rho\right) T.\end{aligned} \tag{31}$$

In addition, the probability that the ω^{th} DU's CB vector is loaded in the other $M\eta-1$ queues is $(M\eta-1)/M$. In this case, the average length over these queues is $(\Omega-Q)/(\eta M-1)$ and, thus, the DU will wait for $(\Omega-Q)M/(\eta M-1) - M/2$ TSs, on average, to be dequeued. Hence, the average latency of the ω^{th} packet dequeued in this case, denoted by $\tau_\omega^{\text{other}}$, is upper bounded by

$$\begin{aligned}\tau_\omega^{\text{other}} &= \tau_\omega^{\text{wait}} + \left(\frac{(\Omega-Q)M}{\eta M-1} - \frac{M}{2}\right) N_oT \\ &\leqslant (\Omega-\omega)\,\beta T + \left[\frac{\omega-\Omega}{\eta} + \left(\frac{\Omega-Q}{\eta 2^{K_o}-1} - \frac{1}{2}\right) 2^{K_o}\right] \\ &\quad \times D^{-1}\left(D(N,\epsilon,\rho) - K_o, \epsilon, \rho\right) T.\end{aligned} \tag{32}$$

As a result, the mean latency for the ω^{th} DU in OB-based short-packet communications is given by

$$\tau_\omega^{\text{O}} = 2^{-K_o}\tau_\omega^{\text{longest}} + (1-\eta)\tau_\omega^{\text{empty}} + (\eta - 2^{-K_o})\tau_\omega^{\text{other}}, \tag{33}$$

where $\tau_\omega^{\text{longest}}$, $\tau_\omega^{\text{empty}}$ and $\tau_\omega^{\text{other}}$ are given in (30), (31) and (32), respectively.

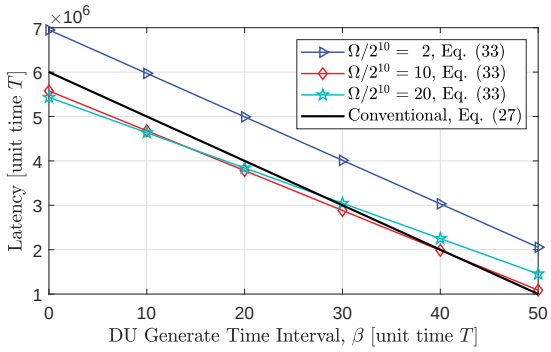


Fig. 20. Mean latency of the 10^5th packet in OB-based short-packet communications versus DU TS βT, for $\rho = 10$dB, $\epsilon = 10^{-3}$, $L = 24$, $N = 60$, $K_o = 10$.

In Fig. 19, we first compare the latency excluding initialisation time, i.e., $\beta = 0$, between OB-based and conventional short-packet communications, to investigate the impact of shorter blocklength and packet reordering on the OB-based design. In Fig. 19(a), the average latencies of the 10^5th packet dequeued from the longest and the empty queues are plotted as a function of the OB size K_o, where the ratio $\Omega/2^{K_o} = 2, 10, 20$. As shown in this figure, the latency in the cases of $\Omega/2^{K_o} = 10$ and $\Omega/2^{K_o} = 20$ when the queues are empty is lower than the conventional latency. The difference between the OB-based latency and the conventional latency gets larger as K_o increases. In Fig. 19(b), the mean latency of the 10^5th packet is depicted under the same condition applied in Fig. 19(a). Herein, the OB-based latency is higher than the conventional in the case of $\Omega/2^{K_o} = 2$. However, as the ratio $\Omega/2^{K_o}$ gets larger, the OB-based latency becomes lower than the conventional latency. Further, the mean latency of OB-based design is reduced with the increase in the OB size K_o or the total packet number Ω, mainly because the payload rate is improved as K_o or Ω increases. This offers an opportunity to further lower the latency for the DUs that carry critical information by mapping their OB vectors onto the empty queues in the M-queue storage.

Then, the latency including initialisation time is investigated in Fig. 20, where the mean latency of the ω^{th} DU is plotted for the OB size $K_o = 10$ and the ratio $\Omega/2^{K_o} = 2, 10, 20$. As shown in this figure, the latency including initialisation time of the OB-based transmission is higher than that of the conventional transmission when the generation of a DU takes a long TS βT. As βT decreases, the OB-based latency gets lower than the conventional in the cases of $\Omega/2^{K_o} = 10$ and $\Omega/2^{K_o} = 20$, even taking into account the initialisation time $\Omega\beta T$.

D. OB Size Selection

We remark that, the OB size K_o is a key parameter in the design of OB-based transmissions. As shown in Figs. 4, 5 and 6, larger K_o directly leads to higher maximal payload rate. Therefore, from the perspective of maximal payload rate, the parameter K_o is preferred to be as large as possible. On the other hand, as shown in Fig. 10, given a limited number of

packets in the M-queue storage, larger K_o results in lower TS load efficiency, because the number of TSs mapped by the OBs, 2^{K_o}, is an exponential growth function of K_o.

Based on the formulations of the maximal payload rate and the TS load efficiency, the above performance analysis provides useful reference tools for the selection of K_o in practical systems. To begin with, the ultimate goal of our OB-based design is to enhance the goodput. As shown in (25), the goodput is jointly affected by the TS load efficiency η given in (23), the maximal payload rate R_o given in (10), and the energy gain given in (24). Figs. 14, 15 and 16 reveals that the OB-based goodput is improved as K_o increases, given the ratio $\Omega/2^{K_o}$. This implies that the OB size K_o in an efficient OB-based design is determined by the minimum between the number of packets to be transmitted and the buffer capacity, i.e., the value Ω. In other words, as long as Ω is given, the parameter K_o can be obtained by the ratio $\Omega/2^{K_o}$ that achieves the target goodput. Ideally, if Ω is arbitrarily large, the goodput can be maximised by setting $K_o = K$. In this case, the whole DU of a packet is mapped onto the TS index, and there is no storage unit required to load CBs in the M-queue storage.

However, as shown in Fig. 19(a), the average latency of a packet loaded in the longest queue gets longer as K_o increases. In practice, this problem can be solved by a smart scheduling algorithm that allocates the highest priority with the strictest latency requirement to the packets loaded in the shortest queue.

As a result, the OB size K_o is selected according to the target goodput and the value Ω, for efficient OB-based transmissions.

VI. Conclusion

In this paper, the OB concept was developed in short-packet communications, and the maximal payload rate was formulated in the finite blocklength regime to evaluate the validity of OB-based short-packets. Moreover, the TS load efficiency was formulated to address the feasibility of OB-based short-packets, concerning unloaded TSs. Based on these two formulations, we derived analytical forms for the resource utilisation efficiency and performance, in metrics of energy gain, goodput and latency of our developed OB-based short-packet communications. Illustrative numerical results on the performance analysis substantiated the advantage achieved by OB-based design over conventional short-packet communications. From the theoretical analysis and numerical results, several important insights were reached to facilitate the system design of OB-based short-packets:

(i) In general, the total number of packets contained in the M-queue storage, Ω, needs to be greater than twice the number of queues in the storage, M. This will result in higher TS load efficiency and, accordingly, lead to higher resource utilisation efficiency and better performance.

(ii) Based on the design in (i), higher resource utilisation efficiency and better performance are achieved with the increase in the number of OBs in a DU and with the decrease in the blocklength of a packet, which implies that the OB concept is particularly beneficial to short-packet communications.

(iii) The mean queue length $\Omega/2^{K_o} = 10$ is a setting we would recommend for the OB-based system design. Concerning the saturation in the performance enhancement with respect to energy gain, goodput and latency, this setting can balance the resource utilisation efficiency and the number of storage units required in the M-queue storage.

References

[1] G. Durisi, T. Koch and P. Popovski, "Toward massive, ultrareliable, and low-latency wireless communication with short packets", *Proc. IEEE*, vol. 104, no. 9, pp. 1711-1726, Sept. 2016.

[2] B. Lee, S. Park, D. J. Love, H. Ji and B. Shim, "Packet structure and receiver design for low latency wireless communications with ultra-short packets", *IEEE Trans. Commun.*, vol. 66, no. 2, pp. 796-807, Feb. 2018.

[3] G. Durisi, T. Koch, J. Östman, Y. Polyanskiy, and W. Yang, "Short-packet communications over multiple-antenna Rayleigh-fading channels", *IEEE Trans. Commun.*, vol. 64, no. 2, pp. 618-629, Feb. 2016.

[4] Y. Gu, H. Chen, Y. Li and B. Vucetic, "Ultra-reliable short-packet communications: half-duplex or full-duplex relaying?", *IEEE Wireless Commun. Lett.*, vol. 7, no. 3, pp. 348-351, June 2018.

[5] X. Sun, S. Yan, N. Yang, Z. Ding, C. Shen and Z. Zhong, "Short-packet downlink transmission with non-orthogonal multiple access", *IEEE Trans. Wireless Commun.*, vol. 17, no. 7, pp. 4550-4564, July 2018.

[6] C. Sun, C. She, C. Yang, T. Q. S. Quek, Y. Li and B. Vucetic, "Optimizing resource allocation in the short blocklength regime for ultra-reliable and low-latency communications", *IEEE Trans. Wireless Commun.*, vol. 18, no. 1, pp. 402-415, Jan. 2019.

[7] G. J. Sutton et al., "Enabling technologies for ultra-reliable and low latency communications: From PHY and MAC layer perspectives", *IEEE Commun. Surveys Tuts.*, vol. 21, no. 3, pp. 2488-2524, thirdquarter 2019.

[8] P. Schulz, et al., "Latency critical IoT applications in 5G: Perspective on the design of radio interface and network architecture", *IEEE Commun. Mag.*, vol. 55, no. 2, pp. 70-78, Feb. 2017.

[9] C. Bockelmann, et al., "Massive machine-type communications in 5G: Physical and MAC-layer solutions", *IEEE Commun. Mag.*, vol. 54, no. 9, pp. 59-65, Sep. 2016.

[10] C. E. Shannon, "A mathematical theory of communication", *The Bell System Technical Journal*, vol. 27, no. 3, pp. 379-423, Jul. 1948.

[11] A. Valembois and M. Fossorier, "Sphere-packing bounds revisited for moderate block lengths", *IEEE Trans. Inf. Theory*, vol. 50, pp. 2998–3014, 2004.

[12] J. Shi and R. Wesel, "A study on universal codes with finite block lengths", *IEEE Trans. Inf. Theory*, vol. 53, pp. 3066–3074, 2007.

[13] G. Wiechman and I. Sason, "An improved sphere-packing bound for finite-length codes over symmetric memoryless channels", *IEEE Trans. Inf. Theory*, vol. 54, pp. 1962–1990, 2008.

[14] Y. Polyanskiy, H. V. Poor, and S. Verdú, "Channel coding rate in the finite blocklength regime", *IEEE Trans. Inf. Theory*, vol. 56, no. 5, pp. 2307–2359, May 2010.

[15] T. Erseghe, "Coding in the finite-blocklength regime: Bounds based on Laplace integrals and their asymptotic approximations", *IEEE Trans. Inf. Theory*, vol. 62, no. 12, pp. 6854-6883, Dec. 2016.

[16] J. Östman, G. Durisi, E. G. Ström, M. C. Coşkun and G. Liva, "Short packets over block-memoryless fading channels: Pilot-assisted or noncoherent transmission?", *IEEE Trans. Commun.*, vol. 67, no. 2, pp. 1521-1536, Feb. 2019.

[17] A. Lancho, J. Östman, G. Durisi, T. Koch and G. Vazquez-Vilar, "Saddlepoint Approximations for Short-Packet Wireless Communications", *IEEE Trans. Wireless Commun.*, vol. 19, no. 7, pp. 4831-4846, Jul. 2020.

[18] R. Devassy, G. Durisi, G. C. Ferrante, O. Simeone and E. Uysal, "Reliable Transmission of Short Packets Through Queues and Noisy Channels Under Latency and Peak-Age Violation Guarantees", *IEEE J. Sel. Areas Commun.*, vol. 37, no. 4, pp. 721-734, Apr. 2019.

[19] D. Zheng, Y. Yang, L. Wei and B. Jiao, "Decode-and-Forward Short-Packet Relaying in the Internet of Things: Timely Status Updates", *IEEE Trans. Wireless Commun.*, doi: 10.1109/TWC.2021.3093163.

[20] A. T. P. Nguyen, R. Le Bidan and F. Guilloud, "Trade-Off Between Frame Synchronization and Channel Decoding for Short Packets," *IEEE Commun. Lett.*, vol. 23, no. 6, pp. 979-982, Jun. 2019.

[21] S. S. Ullah, S. C. Liew, G. Liva and T. Wang, "Short-Packet Physical-Layer Network Coding," *IEEE Trans. Commun.*, vol. 68, no. 2, pp. 737-751, Feb. 2020.

[22] Y. Yang and B. Jiao, "Information-guided channel-hopping for high data rate wireless communication", *IEEE Commun. Lett.*, vol. 12, no. 4, pp. 225–227, Apr. 2008.
[23] N. Ishikawa, S. Sugiura and L. Hanzo, "50 years of permutation, spatial and index modulation: From classic RF to visible light communications and data storage", *IEEE Commun. Surveys Tuts.*, vol. 20, no. 3, pp. 1905-1938, Mar. 2018.
[24] B. Jiao, "A high spectral efficiency method enabled by opportunistic-bit", *IEEE Trans. Veh. Technol.*, vol. 67, no. 12, pp. 12368-12372, Dec. 2018.
[25] Y. Yang, "Permutation-based transmissions in ultra-reliable and low-latency communications", *IEEE Commun. Lett.*, vol. 25, no. 3, pp. 1024-1028, Mar. 2021.
[26] Y. Yang and L. Hanzo, "Permutation-based TCP and UDP transmissions to improve goodput and latency in the Internet-of-Things", *IEEE Internet Things J.*, doi: 10.1109/JIOT.2021.3068238.
[27] M. Yin, J. Chen, B. Jiao, and H. V. Poor, "Utilization of opportunistic-bits with paired transmissions", *IEEE Wireless Commun. Lett.*, vol. 8, no. 3, pp. 661-664, Jun. 2019.
[28] C. Bormann, C. Burmeister and M. Degermark, Robust header compression (ROHC): Framework and four profiles: RTP UDP ESP and uncompressed, Fremont, CA, USA, Jul. 2001.
[29] S. Salamat Ullah, S. C. Liew, G. Liva and T. Wang, "Short-Packet Physical-Layer Network Coding", *IEEE Trans. Commun.*, vol. 68, no. 2, pp. 737-751, Feb. 2020.
[30] J. Wu, W. Kim and B. Shim, "Pilot-Less One-Shot Sparse Coding for Short Packet-Based Machine-Type Communications", *IEEE Trans. Veh. Technol.*, vol. 69, no. 8, pp. 9117-9120, Aug. 2020.
[31] L. Becchetti, A. Clementi, E. Natale, et al., "Self-stabilizing repeated balls-into-bins", *Distrib. Comput.*, vol. 32, no. 1, pp. 59–68, Feb. 2019.
[32] N. Cancrini, G. Posta, "Propagation of chaos for a balls into bins model", *Electronic Communications in Probability*, vol. 24, no. 1, pp. 1-9, Jan. 2019.
[33] M. Abramowitz and I. A. Stegun, Handbook of Mathematical Functions, New York, NY, USA:Dover, 1972.
[34] A. Berman and R.J. Piemmons, Nonnegative Matrices in the Mathematical Sciences, PA, Philadelphia:SIAM, 1994.

Mingxi Yin received the B.Eng. degree in electronics engineering from the Beijing University of Posts and Telecommunications (BUPT), Beijing, China, in 2016, and the Ph.D. degree from the Department of Electronics, Peking University, Beijing, China, in 2021. Her research interests include wireless communications in the physical layer and wireless digital health.

Yuli Yang (S'04-M'08-SM'19) received her Ph.D. degree in Communications and Information Systems from Peking University in July 2007. Since Dec 2019, she has been with the University of Lincoln as a Senior Lecturer in Electrical/Electronic Engineering. From Jan 2010 to Dec 2019, she was with King Abdullah University of Science & Technology, Melikşah University, and the University of Chester on various academic positions. Her industry experience includes working as a Research Scientist with Bell Labs Shanghai, from Aug 2007 to Dec 2009, and an Intern Researcher with Huawei Technologies, from June 2006 to July 2007. Her research interests include modelling, design, analysis and optimization of wireless systems and networks.

Jen-Ming Wu received the B.Sc. degree from National Taiwan University, Taipei, Taiwan, and Ph.D. from University of Southern California, all in Electrical Engineering. From '98 –'03, he has been with Sun Microsystems Inc. in Sunnyvale, CA, USA as Member of Technical Staff. Since 2003, he has been with the faculty of Inst. of Communications Engineering, Dept. of Electrical Engineering, National Tsing Hua University, Taiwan, where he currently holds Full Professor position. The research interests of Prof. Wu have covered a range of areas in communications from theory to practice, including wireless communications signal processing, information and coding theory, MIMO radar signal processing, communications transceiver IC designs, and wireless system prototypes. He holds more than 8 US patents in the field of communications, 14 contributions in 3GPP 5G New Radio and IEEE 802.16m standards meetings, and has published more than 100 technical articles in international journals and conference proceedings. He has received the Best Paper Award of Taiwan Telecommunications Symposium (2008 & 2020), the Outstanding Chapter Award of IEEE Vehicular Technology Society Taipei Chapter (2020), the Achievement Award of the Golden Silicon Award Competition (2007 & 2008), and the Bronze Medal of Taiwan Semiconductor Manufacture Co-operation(TSMC) Research Program (2007). He has served as the Chair of IEEE Vehicular Technology Society Taipei Chapter, the TPC chair of IEEE Asia Pacific Wireless Communications Symposium (2018), the TPC member of IEEE Global Communications Conference (Globecom), IEEE International Communications Conference (ICC), IEEE Vehicular Technology Conference (VTC), and many other conferences.

Bingli Jiao (M'05-SM'11) received the B.S. and M.S. degrees from Peking University, China, in 1983 and 1988, respectively, and the Ph.D. degree from Saarland University, Germany, in 1995. He became an Associate Professor in 1995 and a Professor with Peking University in 2000. He currently is the director of Wireless Communication and Signal Processing Research Center, Peking University, Beijing, China. He is also a director of the Joint Laboratory for Advanced Communication Research between Peking University and Princeton University. His current research interests include full-duplex communications, information theory, and signal processing. He is a pioneer of co-frequency and co-time full-duplex as found in his early patent in 2006.